# An Analysis of Criminal Justice Data

*Sam Loyd*

*March 2, 2019*

## Background

Several years ago, one of my son's friends was abducted and assaulted by her school bus driver. My son who was only 15 at the time worked with local law enforcement to help track her down by contacting common friends with social media, his cell phone and various apps at his disposal. Since then, he has been fascinated with criminal justice and in pursuing a career in that field. He has just completed his first semester of study in criminal justice in college. Of course, that makes his mom and I a bit nervous given the media portrayal of how dangerous a career in that field is.

## Interests

I am trying to address forming a realistic level of concern around hazards in my son's chosen profession in criminal justice in addition to exploring a few areas of interests that I have in relation to race, region and gender using criminal justice data. I plan to address this by fully researching the three data sets that I have found so far. I have two key interests. The first is to make recommendations to my son that will minimize my concerns about the dangers that he may face. The second is to assess how gender and/or race might impact arrests and ultimately police shootings.

## Problem Statement

Evaluate the three datasets obtained from the Washington Post and the FBI to gain a realistic understanding of the dangers faced by law enforcement and to use the same data to evaluate the impact of possible bias by law enforcement in relation to race and gender that might impact interactions with the public as measured through arrests and police shootings.

## Research Questions

### A. Safety Concerns

- Should I encourage my son to move to a new area in the country based on real evidence from the data?
- What type of community would be the safest working environment for him?
- Is there a safer shift that he should consider?

### B. General Interests

- Is there a correlation between gender and the total number of arrests?
- Is there a correlation between race and ethnicity categories and the total number of arrests?
- Is there a correlation between race and police shootings?
- The class asked if there is a correlation between arrests of residents and undocumented Hispanic immigrants?

## Audience

Given my interest and a broader exploration of variables that I am already considering for the work, the audience will be this class which by the end of the semester should allow for a technical exploration. I want to feel free to use all the tools that I will learn to explore the data. I was tempted to use the general public as my audience, but I felt that doing so might limit the scope at worst and require a lot of explaining at best, and I wanted to focus more on what the data had to say.

# Purpose

Answers to these questions will help me talk to my son with a level of confidence about my concerns for his safety. It will also allow me to inform the class what I learn about working hazards for law enforcement workers in addition to concerns about how their bias might affect the public at large.

# Methodology:

1. Data Acquisition

- Kaggle
- FBI

2. Data Wrangling

- Missingness Analysis (MAR)
- New Summaries

3. Data analysis - EDA

- Distribution Analysis
- Correlation Analysis

4. Modeling

5. Reporting

# Data Sets

## Washington Post: Fatal Shootings By Police Data Set

The first data set which piqued my interest when I randomly started looking for datasets was titled, The Washington Post's Dataset of Fatal Police Shootings in the US since 2015 available from the following hyperlink on kaggle.com.

https://www.kaggle.com/brendanhasz/fatal-police-shootings#fatal-police-shootings-data.csv

Acknowledgements and licenses for the data can be found at the following hyperlink. There was no codebook, but a description of the variables is available here as well.

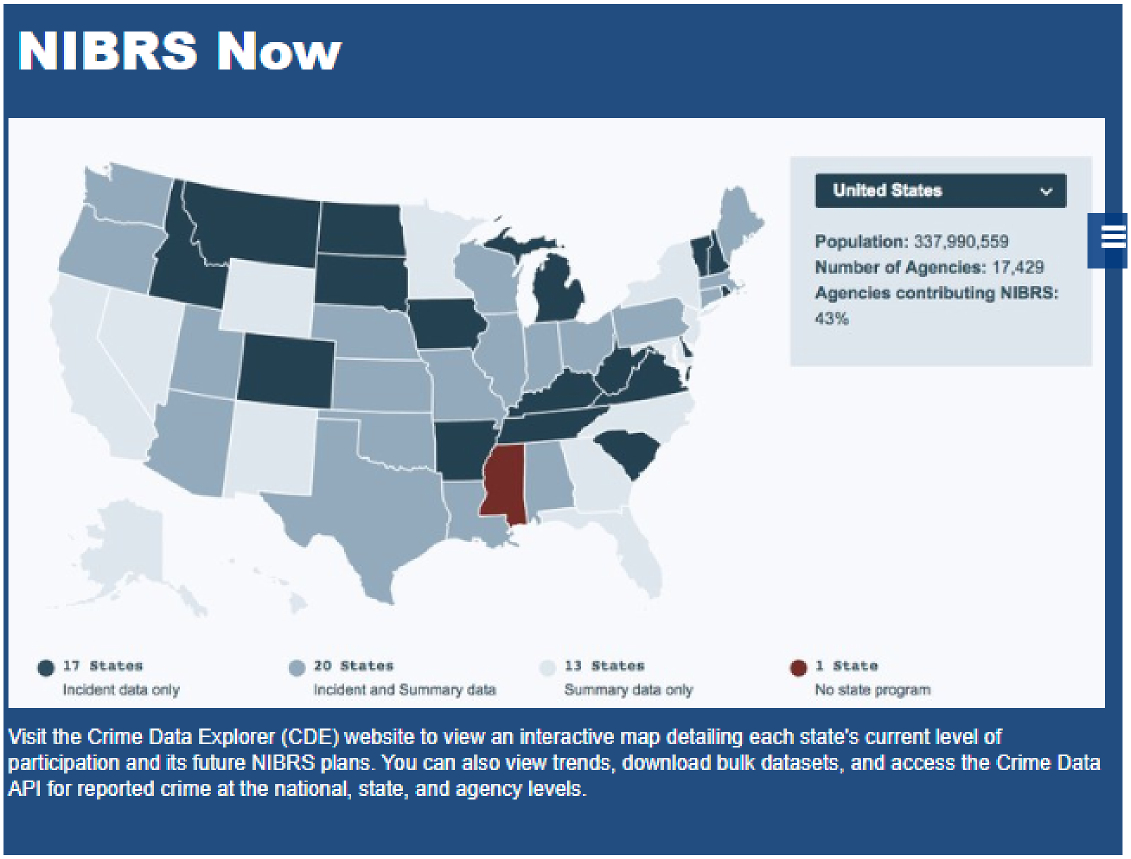
https://www.kaggle.com/brendanhasz/fatal-police-shootings/home

The data was provided by Brendan Has, a PHD student with the University of Minnesota. There were 14 variables in the original dataset. The data in this set included categorical variables such as id, name of individual shot, date of shooting, race, gender, city, state, evidence of mental illness, flight during incident, if the individual was armed, threat level and whether the officer involved wore a body camera. A count will be used summing various categorical values in my evaluation, and I have done early work on age which is quantitative. Race values were missing for several of the incidents reported as well as missing data for age and a weapon if armed. Missing data is inconsistent in this data and ranges from a NA value to just an empty column. There was one observation recorded per fatal shooting.

*Fatal Policie Shooting Data Set Structure:*

```
## Observations: 1,958
## Variables: 14
## $ id                      <int> 3, 4, 5, 8, 9, 11, 13, 15, 16, 17, 19,...
## $ name                    <fct> Tim Elliot, Lewis Lee Lembke, John Pau...
## $ date                    <date> 2015-01-02, 2015-01-02, 2015-01-03, 2...
## $ manner_of_death         <fct> shot, shot, shot and Tasered, shot, sh...
## $ armed                   <fct> gun, gun, unarmed, toy weapon, nail gu...
## $ age                     <int> 53, 47, 23, 32, 39, 18, 22, 35, 34, 47...
## $ gender                  <fct> M, M, M, M, M, M, M, M, F, M, M, M, M,...
## $ race                    <fct> A, W, H, W, H, W, H, W, W, B, W, B, B,...
## $ city                    <fct> Shelton, Aloha, Wichita, San Francisco...
## $ state                   <fct> WA, OR, KS, CA, CO, OK, AZ, KS, IA, PA...
## $ signs_of_mental_illness <fct> True, False, False, True, False, False...
## $ threat_level            <fct> attack, attack, other, attack, attack,...
## $ flee                    <fct> Not fleeing, Not fleeing, Not fleeing,...
## $ body_camera             <fct> False, False, False, False, False, Fal...
```

### NIBRS Data Sets



NIBRS Now

Population: 337,990,559
Number of Agencies: 17,429
Agencies contributing NIBRS: 43%

Visit the Crime Data Explorer (CDE) website to view an interactive map detailing each state's current level of participation and its future NIBRS plans. You can also view trends, download bulk datasets, and access the Crime Data API for reported crime at the national, state, and agency levels.

#### NIBR National Arrests Data Set

From this I wanted to compare certain frequencies against overall interactions with the public that involved arrests through the Arrest Data - Reported Number of Adult Arrests by Crime dataset available from the FBI's Crime Data Explorer. The dataset was recorded by the FBI for years 1995-2016 and was last modified January 1, 2017. The data for a given year only includes agencies that reported for the full 12 months.

The dataset has 84 observations and 11 variables. This data set also includes the year and offense which are categorical and totals for gender and race which are quantitative. There is one observation per year per type of offense.

This dataset can be found at the following link.

https://crime-data-explorer.fr.cloud.gov/downloads-and-docs

*NIBR National Arrest Data Set Structure:*

```
## Observations: 84
## Variables: 11
## $ year                    <fct> 2016, 2016, 2016, 2016, 2016, 2016, 201...
## $ offense_code            <fct> ASR_ARSON, ASR_AST, ASR_AST_SMP, ASR_BR...
## $ offense_name            <fct> Arson, Aggravated Assault, Simple Assau...
## $ population              <int> 264534532, 264534532, 264534532, 264534...
## $ total_male              <int> 4509, 224176, 570193, 116213, 180722, 9...
## $ total_female            <int> 1426, 67016, 213178, 28754, 68577, 2847...
## $ race_population         <int> 263887632, 263887632, 263887632, 263887...
## $ white                   <int> 4263, 183478, 514297, 101778, 161655, 8...
## $ black                   <int> 1373, 94982, 237138, 39235, 73552, 3258...
## $ asian_pacific_islander <int> 103, 5365, 12418, 2035, 2556, 14813, 34...
## $ american_indian         <int> 183, 6129, 14376, 1323, 9460, 11743, 23...
```

In order to answer the question posed from the class, I also downloaded the datasets used to build the Arrest Data - Reported Number of Adult Arrests by Crime dataset above. They were obtained from the same website and data source and were merged for years 2014 through 2016 into a data set that will be referenced as the National Incident Based Records or NIBR data set. This data held variables for ethnicity and residency status which could help with a question from the class.

### NIBR Daily Arrest Data Set

*NIBR Daily Arrest Total Data Set Structure:*

```
## Observations: 1,096
## Variables: 13
## $ Arrest_Date  <date> 2014-01-01, 2014-01-02, 2014-01-03, 2014-01-04, ...
## $ Caucasion    <int> 2594, 1935, 2215, 2261, 1909, 1816, 1988, 2118, 2...
## $ Black        <int> 1030, 748, 876, 940, 711, 674, 802, 910, 1028, 10...
## $ Native       <int> 68, 44, 59, 49, 52, 42, 44, 44, 54, 58, 59, 51, 6...
## $ Asian        <int> 27, 18, 23, 35, 33, 25, 16, 33, 29, 36, 25, 19, 2...
## $ Hispanic     <int> 332, 224, 235, 292, 226, 212, 236, 228, 293, 318,...
## $ Non_Hispanic <int> 2775, 2078, 2398, 2429, 2007, 1932, 2180, 2338, 2...
## $ Non_Resident <int> 803, 686, 757, 823, 665, 567, 678, 709, 888, 923,...
## $ Resident     <int> 2468, 1722, 2034, 2039, 1724, 1664, 1844, 2035, 2...
## $ Undocumented <int> 72, 51, 61, 69, 55, 52, 53, 52, 72, 76, 71, 57, 6...
## $ Male         <int> 2802, 1976, 2260, 2357, 1943, 1820, 2057, 2244, 2...
## $ Female       <int> 974, 798, 958, 980, 797, 773, 838, 898, 1106, 116...
## $ Arrest_Total <int> 3776, 2774, 3218, 3337, 2740, 2593, 2895, 3142, 3...
```

## Officer Injury and Fatality Data Set

And finally, also from the same FBI data source, I chose to look at assaults against police officers in the Assaults on Law Enforcement Officers dataset which can be found at the same link above. This data was last updated on September 24, 2018 for years 1995-2017. It has 31 variables. I will not use all the variables, but of key interest to me are the fatality and injury related columns. I am currently planning on using year, region, population group, and various injury and fatality counts taken from this data set. Year, region and population group are all categorical while the two-hour time ranges of assault and injury variables that both provide totals are quantitative. There was one observation recorded per year per reporting agency.

*Summarized Officer Injury Data Set Structure:*

```
## Observations: 43,663
## Variables: 30
## $ DATA_YEAR                      <fct> 2014, 2014, 2014, 2014, 2014, 2...
## $ PUB_AGENCY_NAME                <fct> 15th Circuit Drug Enforcement U...
## $ STATE_ABBR                     <fct> SC, TN, GA, LA, SC, SC, ID, MD,...
## $ DIVISION_NAME                  <fct> South Atlantic, East South Cent...
## $ REGION_NAME                    <fct> South, South, South, South, Sou...
## $ AGENCY_TYPE_NAME               <fct> Other, Other, City, City, City,...
## $ POPULATION_GROUP_DESC          <fct> Cities under 2,500, Cities unde...
## $ COUNTY_NAME                    <fct> HORRY, ANDERSON, WILCOX, VERMIL...
## $ TIME_0001_0200_CNT             <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 5...
## $ TIME_0201_0400_CNT             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3...
## $ TIME_0401_0600_CNT             <int> 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0...
## $ TIME_0601_0800_CNT             <int> 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0...
## $ TIME_0801_1000_CNT             <int> 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0...
## $ TIME_1001_1200_CNT             <int> 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 2...
## $ TIME_1201_1400_CNT             <int> 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0...
## $ TIME_1401_1600_CNT             <int> 0, 0, 0, 4, 0, 0, 0, 1, 0, 1, 0...
## $ TIME_1601_1800_CNT             <int> 0, 0, 0, 3, 0, 0, 0, 0, 0, 1, 0...
## $ TIME_1801_2000_CNT             <int> 0, 0, 0, 7, 0, 0, 0, 0, 0, 1, 1...
## $ TIME_2001_2200_CNT             <int> 0, 0, 0, 7, 0, 0, 0, 0, 0, 1, 0...
## $ TIME_2201_0000_CNT             <int> 0, 0, 0, 2, 0, 0, 0, 0, 7, 0, 2...
## $ FIREARM_INJURY_CNT             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ FIREARM_NO_INJURY_CNT          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ KNIFE_INJURY_CNT               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ KNIFE_NO_INJURY_CNT            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ HANDS_FISTS_FEET_INJURY_CNT    <int> 0, 0, 0, 3, 0, 0, 0, 0, 7, 0, 6...
## $ HANDS_FISTS_FEET_NO_INJURY_CNT <int> 0, 0, 0, 1, 0, 0, 0, 1, 0, 4, 6...
## $ OTHER_INJURY_CNT               <int> 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0...
## $ OTHER_NO_INJURY_CNT            <int> 0, 0, 0, 34, 0, 0, 0, 0, 0, 0, ...
## $ LEOKA_FELONY_KILLED            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ LEOKA_ACCIDENT_KILLED          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

# Data Wrangling

## Police Shooting Data Set

The original data set was by individual shootings. I decided to summarzie those by date to allow for better analysis.

*Summarized By Day Police Shootin Data Set Structure:*

```
## Observations: 680
## Variables: 6
## $ Shooting_Date  <date> 2015-01-02, 2015-01-03, 2015-01-04, 2015-01-05...
## $ Caucasion      <dbl> 1, 0, 2, 0, 3, 1, 3, 1, 1, 2, 4, 2, 1, 0, 2, 1,...
## $ Black          <dbl> 0, 0, 0, 0, 1, 3, 1, 0, 0, 0, 2, 2, 0, 1, 0, 0,...
## $ Asian          <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Hispanic       <dbl> 0, 1, 1, 1, 0, 0, 0, 1, 2, 0, 0, 1, 0, 2, 0, 0,...
## $ Fatality_Total <dbl> 2, 1, 3, 1, 4, 4, 4, 2, 3, 2, 6, 5, 2, 3, 2, 1,...
```

## NIBR Data Sets

### National Arrest Data Set

This data set was already summarized to a degree that did not allow for any significant wrangling.

### NIBR Daily Arrest Data Set

As mentioned above, the original NIBR National Arrest total data set was too summarized to adequately address several questions. I had to go back to the original NIBR website and pull from the states that provided detailed data. I then had to summarize that data in a format to allow for proper correlation analysis. I chose to summarize this data by date using totals for race groupings and ethnicity groupings in addition to residency.

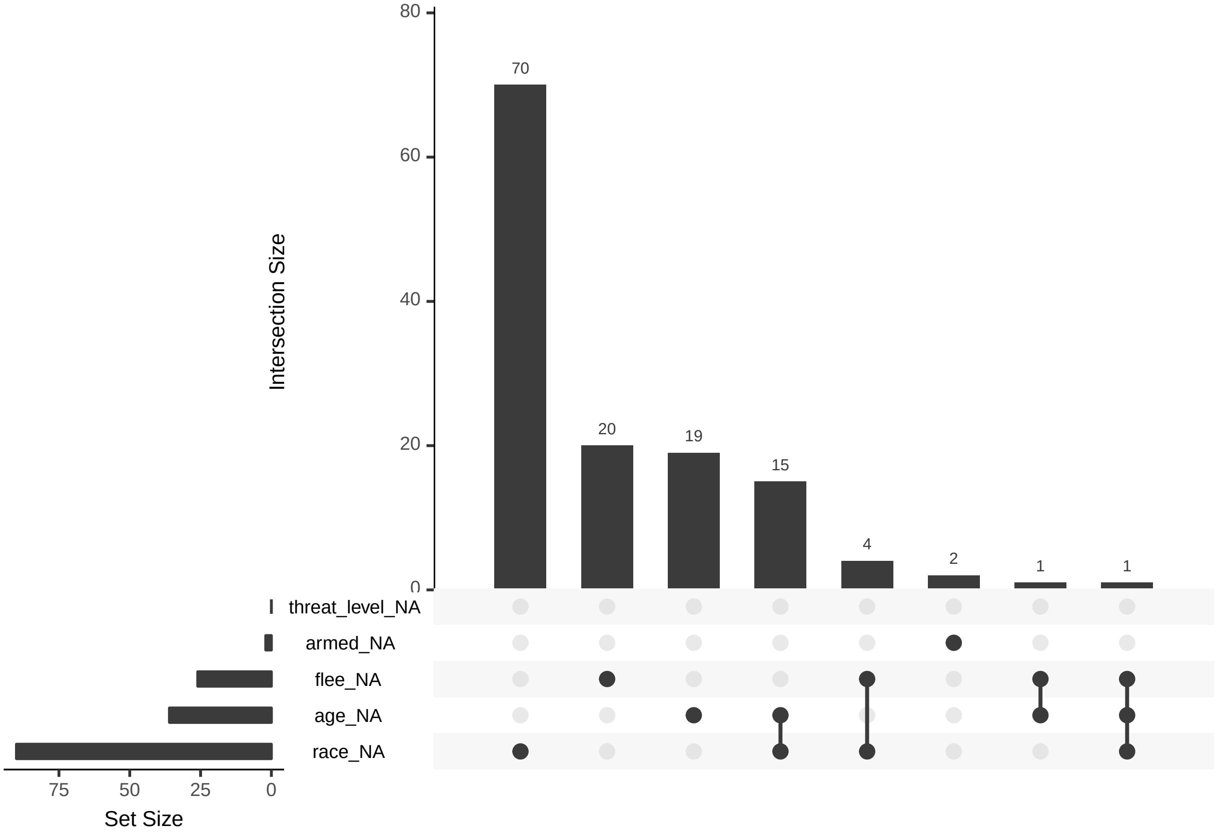
## Officer Injury and Fatality Data Set

No further summarization was required.

# Current Missingness Data Analysis:

## Fatal Police Shootings:

### Original Dataset



*The graph's y-axis is in individual units.*

This graph show overlapping missingness in addition to giving total overall missingness. This would indicate mostly MAR or Missing at Random.

```
# A tibble: 6 x 3
  variable       n_miss pct_miss
  <chr>           <int>    <dbl>
1 Shooting_Date       0        0
2 Caucasion           0        0
3 Black               0        0
4 Asian               0        0
5 Hispanic            0        0
6 Fatality_Total      0        0
```

### Summary NIBR

```
# A tibble: 13 x 3
   variable     n_miss pct_miss
   <chr>         <int>    <dbl>
 1 Arrest_Date       0        0
 2 Caucasion         0        0
 3 Black             0        0
 4 Native            0        0
 5 Asian             0        0
 6 Hispanic          0        0
 7 Non_Hispanic      0        0
 8 Non_Resident      0        0
 9 Resident          0        0
10 Undocumented      0        0
11 Male              0        0
12 Female            0        0
13 Arrest_Total      0        0
```

Due to the summary nature of this data set there is no missingness to discuss.

## Adult Arrests by Crime:

```
# A tibble: 11 x 3
   variable               n_miss pct_miss
   <chr>                   <int>    <dbl>
 1 year                        0        0
 2 offense_code                0        0
 3 offense_name                0        0
 4 population                  0        0
 5 total_male                  0        0
 6 total_female                0        0
 7 race_population             0        0
 8 white                       0        0
 9 black                       0        0
10 asian_pacific_islander      0        0
11 american_indian             0        0
```

Due to the summary nature of this data set there is no missingness to discuss.

## Officer Injury and Fatality

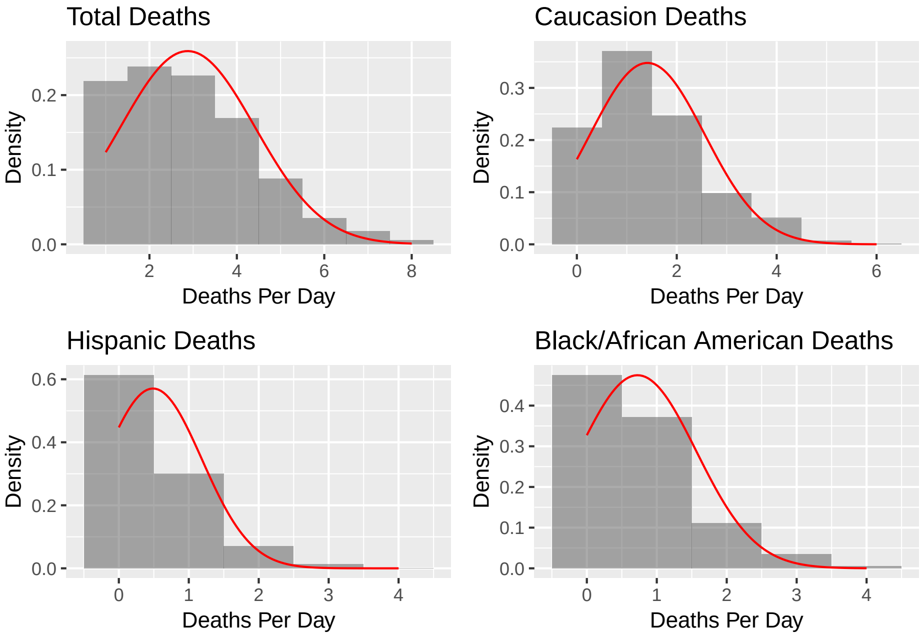
```
# A tibble: 30 x 3
   variable              n_miss pct_miss
   <chr>                  <int>    <dbl>
 1 COUNTY_NAME              813     1.86
 2 DATA_YEAR                0     0
 3 PUB_AGENCY_NAME          0     0
 4 STATE_ABBR               0     0
 5 DIVISION_NAME            0     0
 6 REGION_NAME              0     0
 7 AGENCY_TYPE_NAME         0     0
 8 POPULATION_GROUP_DESC    0     0
 9 TIME_0001_0200_CNT       0     0
10 TIME_0201_0400_CNT       0     0
# ... with 20 more rows
```

A signficant number of counties are missing, but these were not needed for the analysis that is required to answer my qestions.

# Exploratory Data Analysis

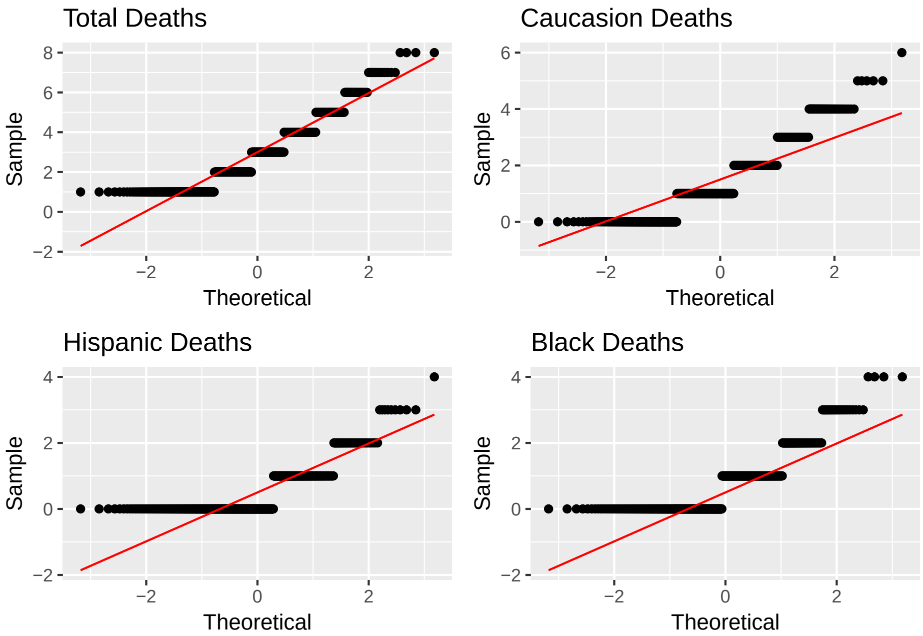
## Police Shooting Data Set

### Distribution Histograms



*These histograms show individual deaths as a unit of measure.*

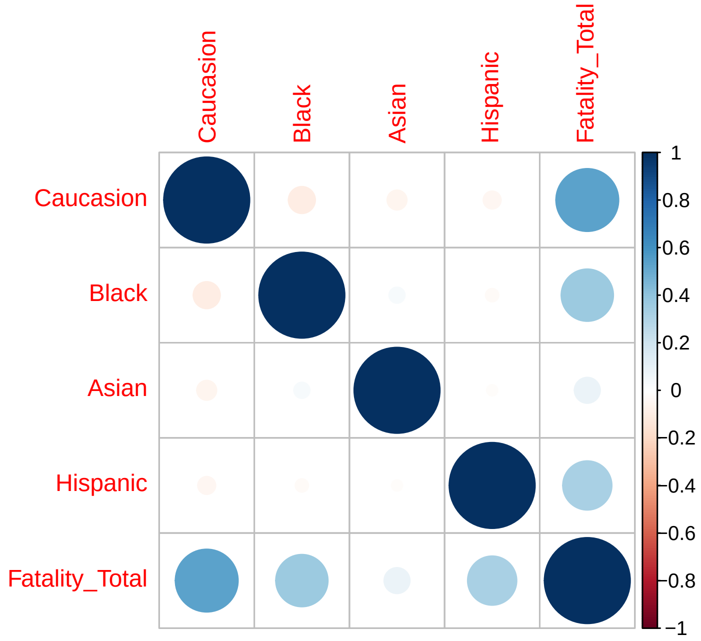
### Probability Plots



### Stat.desc Function

```
##                Caucasion Black Asian Hispanic
## median              1.00  1.00  0.00     0.00
## mean                1.41  0.72  0.04     0.49
## SE.mean             0.04  0.03  0.01     0.03
## CI.mean.0.95        0.09  0.06  0.02     0.05
## var                 1.31  0.71  0.05     0.49
## std.dev             1.15  0.84  0.22     0.70
## coef.var            0.81  1.16  5.07     1.43
## skewness            0.77  1.15  5.36     1.42
## skew.2SE            4.12  6.12 28.60     7.59
## kurtosis            0.32  1.11 30.84     1.92
## kurt.2SE            0.85  2.97 82.39     5.12
## normtest.W          0.88  0.77  0.19     0.69
## normtest.p          0.00  0.00  0.00     0.00
```

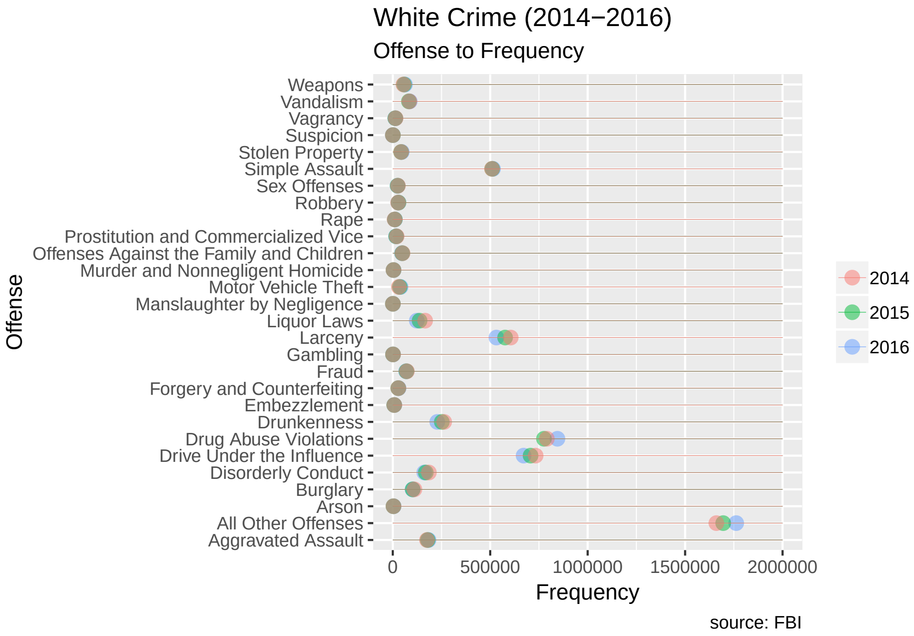
The graphs indicate a non normal distribution with high ranking. For this reason I will use the Kendall method for correlation analysis.



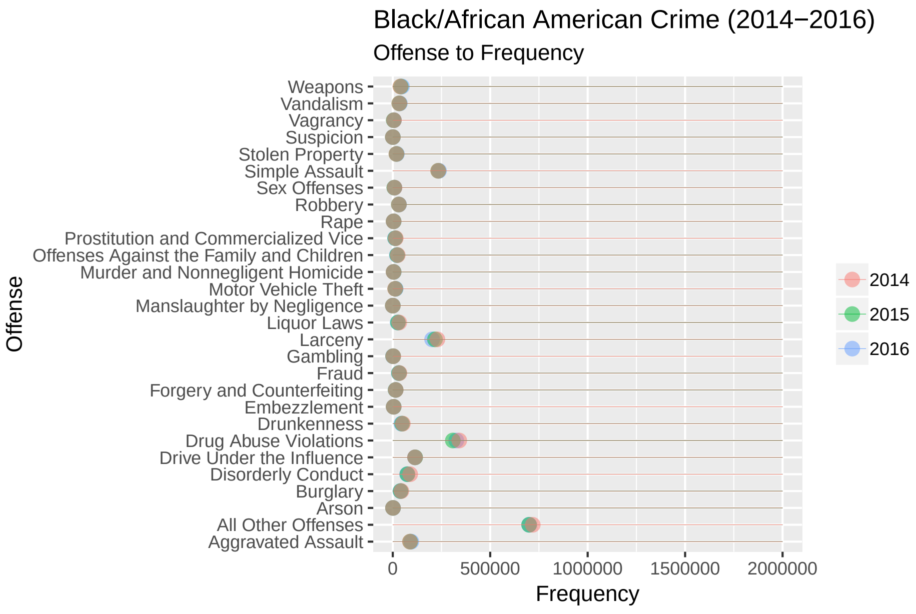
The graph shows that of the racial totals, the one for caucausians seems to have the highest correlation to total fatalities.

## NIBR Data Sets

### NIBR National Arrest Data Set



*X-Axis shows individual offenses as a unit of measure.*

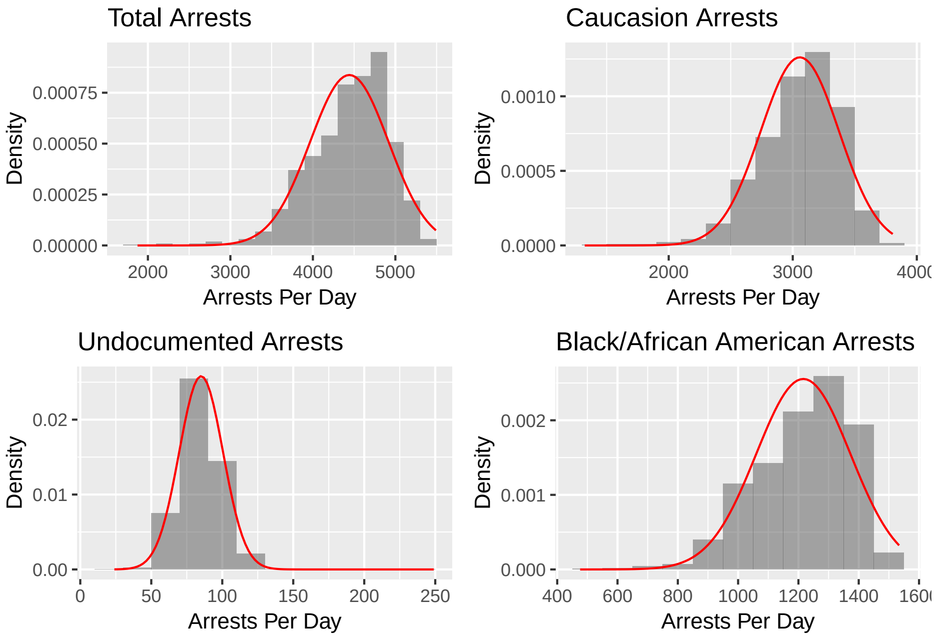


source: FBI

*X-Axis shows individual offenses as a unit of measure.*

### NIBR Daily Arrest Data Set

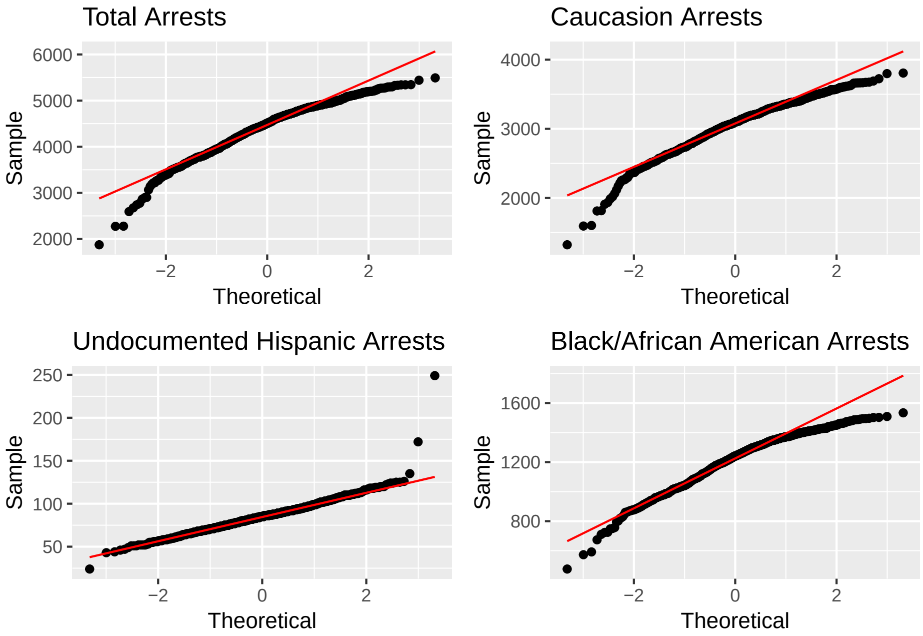
### Distribution Histograms



*Histogram shows individual arrests as a unit of measure.*

In order to perform correlation analysis on this data, I needed to know if it was normally distributed to ascertain which correlation model to apply. It is readily apparent that all but undocumented arrests are skewed to the left and undcoumented is right skewed. This does not appear to be a normally distributed data set. This would lend to using non parametric correlation analysis such as Spearman.

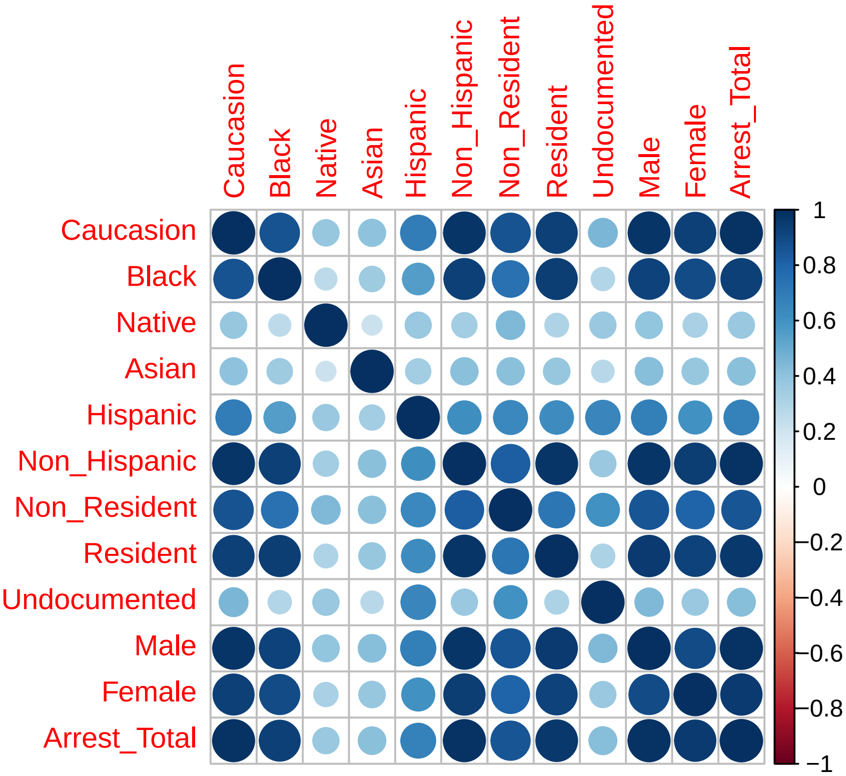
### Distribution Probability Plots



### Stat.desc Function

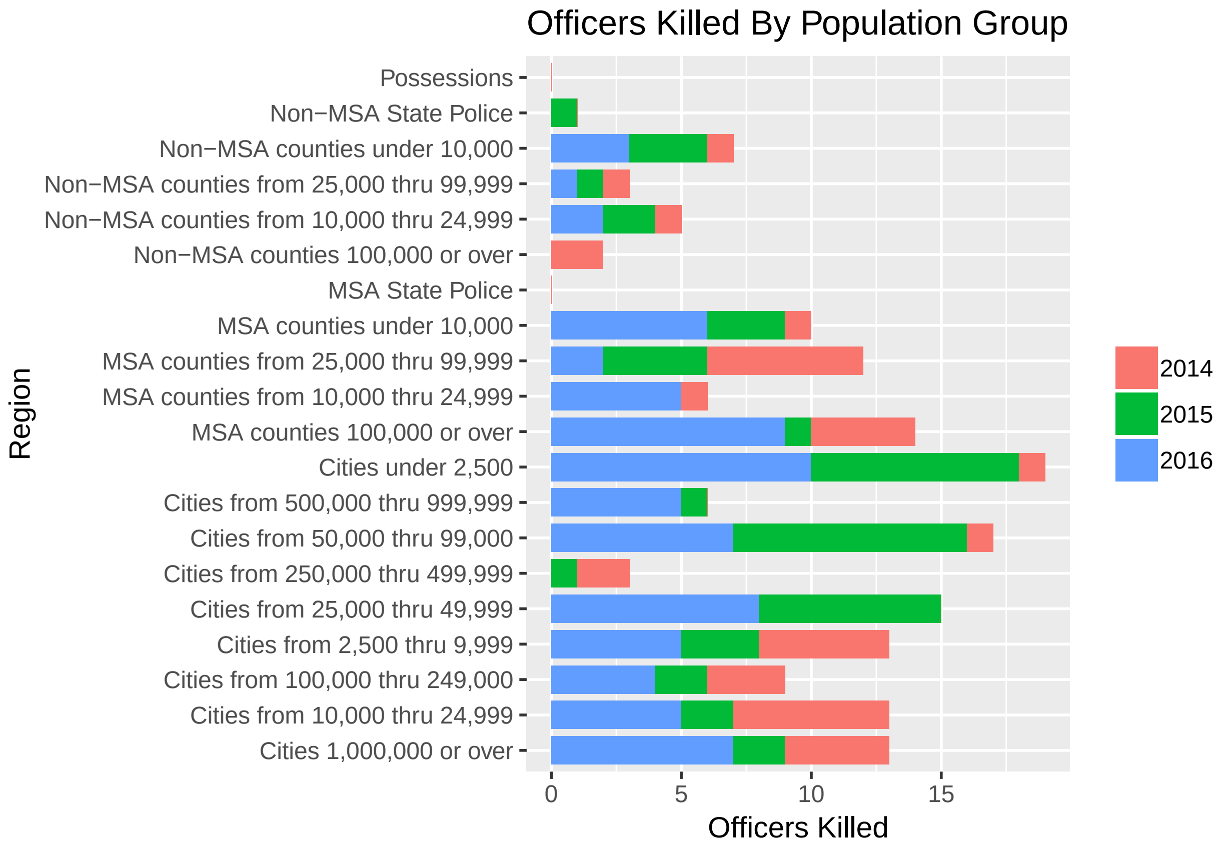
```
##              Caucasion    Black Native Asian Hispanic
## median         3097.50  1242.00  62.00 37.00   356.00
## mean           3057.14  1216.29  62.51 37.96   355.15
## SE.mean           9.56     4.72   0.38  0.25     1.16
## CI.mean.0.95     18.75     9.26   0.75  0.48     2.27
## var          100068.85 24394.81 158.04 66.10  1463.00
## std.dev         316.34   156.19  12.57  8.13    38.25
## coef.var          0.10     0.13   0.20  0.21     0.11
## skewness         -0.85    -0.72   3.58  0.47    -0.22
## skew.2SE         -5.72    -4.86  24.24  3.19    -1.46
## kurtosis          1.69     0.49  52.50  0.64     1.16
## kurt.2SE          5.72     1.65 177.78  2.16     3.94
## normtest.W        0.96     0.96   0.85  0.99     0.99
## normtest.p        0.00     0.00   0.00  0.00     0.00

##              Non_Hispanic Non_Resident Resident Undocumented
## median            3298.00      1112.00  2756.00        85.00
## mean              3242.80      1094.65  2715.77        85.10
## SE.mean             11.13         4.09     8.81         0.47
## CI.mean.0.95        21.84         8.03    17.28         0.92
## var             135783.62     18344.61 85035.15       238.99
## std.dev            368.49       135.44   291.61        15.46
## coef.var             0.11         0.12     0.11         0.18
## skewness            -0.81        -0.70    -0.76         1.26
## skew.2SE            -5.51        -4.76    -5.12         8.54
## kurtosis             1.04         1.53     0.88        11.89
## kurt.2SE             3.52         5.19     2.97        40.27
## normtest.W           0.96         0.97     0.96         0.94
## normtest.p           0.00         0.00     0.00         0.00
##                       Male    Female Arrest_Total
## median            3172.00   1338.00      4508.00
## mean              3134.60   1307.48      4442.08
## SE.mean              9.82      4.86        14.40
## CI.mean.0.95        19.26      9.53        28.25
## var             105653.42 25865.46    227204.02
## std.dev            325.04    160.83       476.66
## coef.var             0.10      0.12         0.11
## skewness            -0.84     -0.80        -0.87
## skew.2SE            -5.67     -5.39        -5.90
## kurtosis             1.52      1.03         1.46
## kurt.2SE             5.15      3.47         4.95
## normtest.W           0.96      0.96         0.96
## normtest.p           0.00      0.00         0.00
```

The notable skew and graphs all support a non normal distribution. There does not appear to be significant ranking. Due to this, I have chosen Spearman's model for correlation testing.

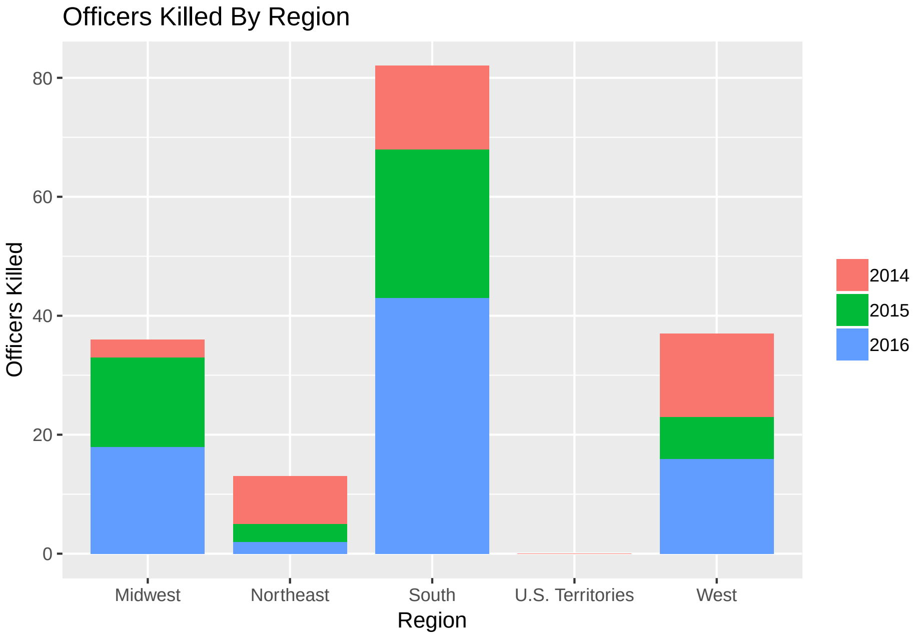


The correlation graph can be used as a matrix to determine the degree to which two variables are correlated in the model observed. In this case the Spearman method was used with the cor function in R to determine correlation. Shading and Size show signficance while color shows positive or negative correlation. In this case all were positive.

## Officer Injury and Fatality Data Set

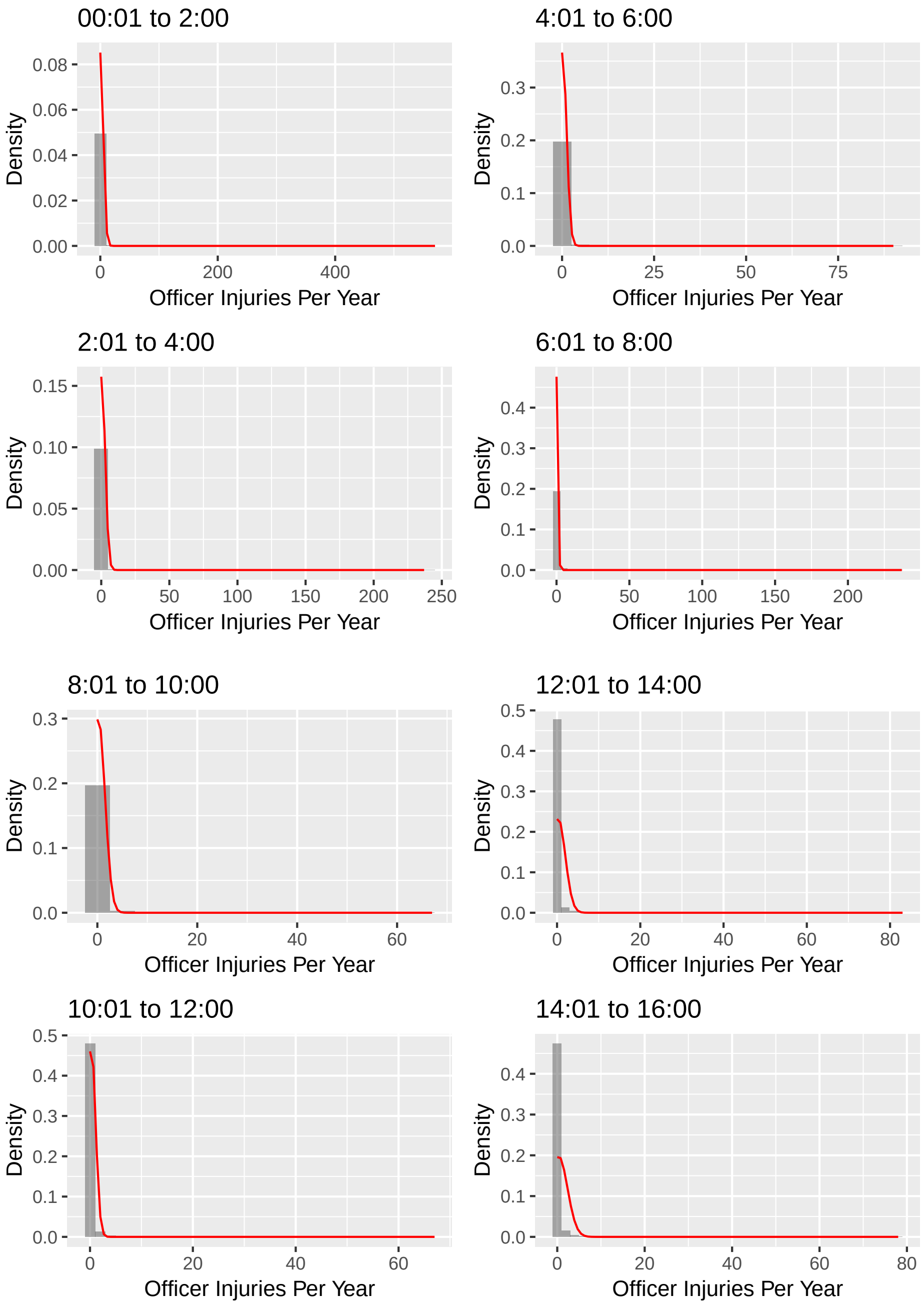


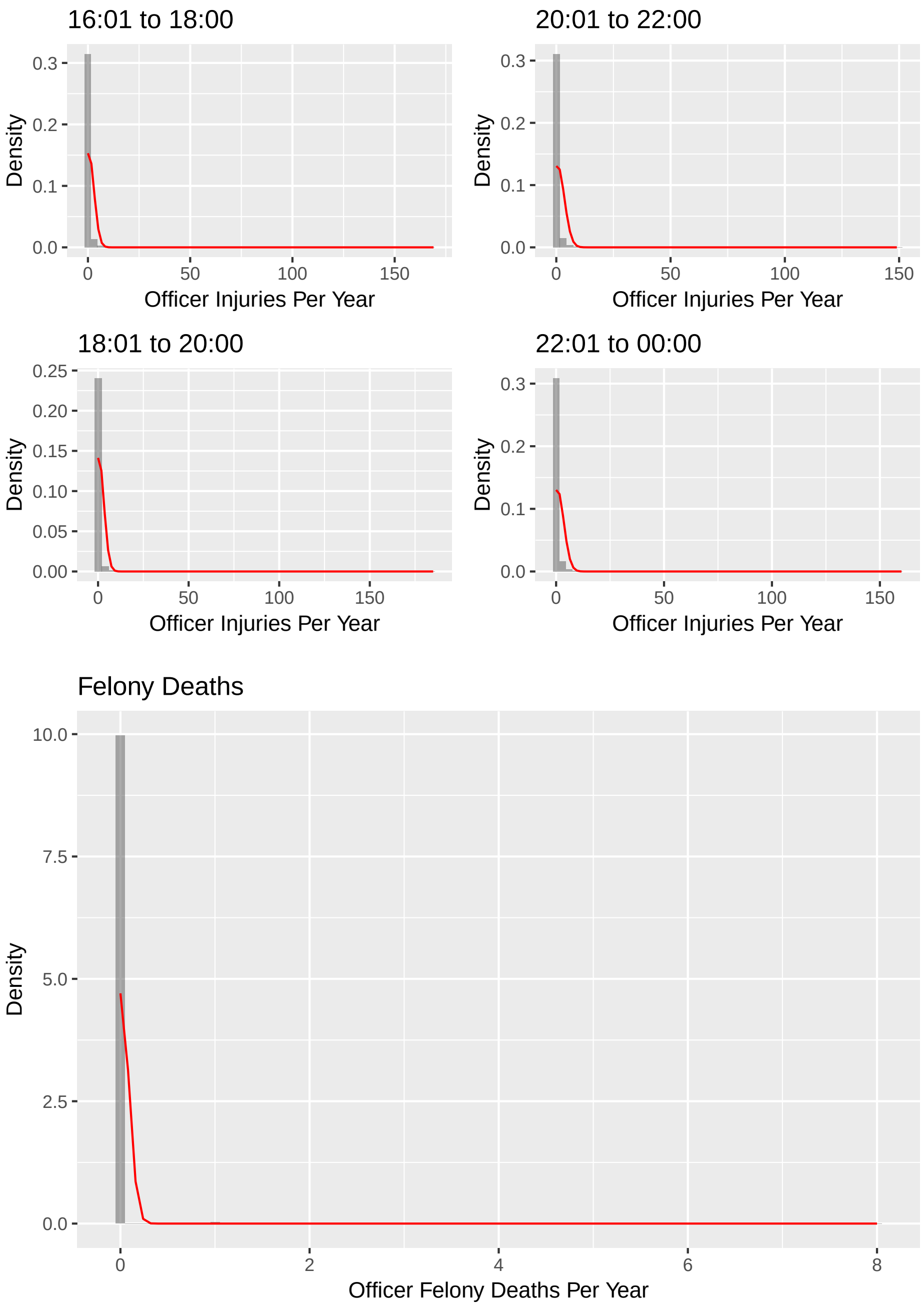
*Graph is shown in individual deaths.*



*Graph is shown in individual deaths.*

### Officer Injury and Fatality Distribution Histograms





16:01 to 18:00

20:01 to 22:00

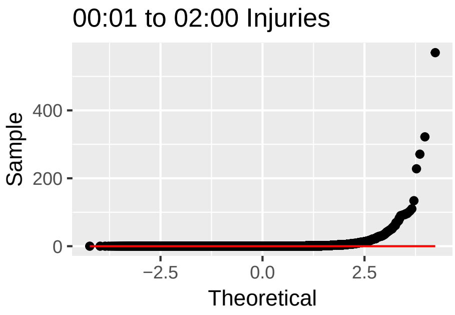
18:01 to 20:00

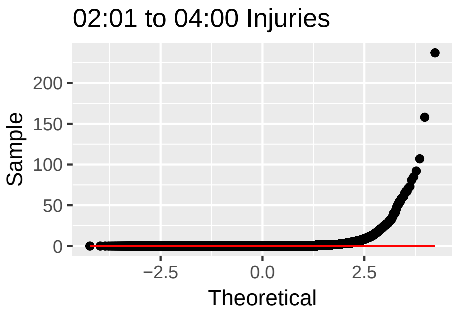
22:01 to 00:00

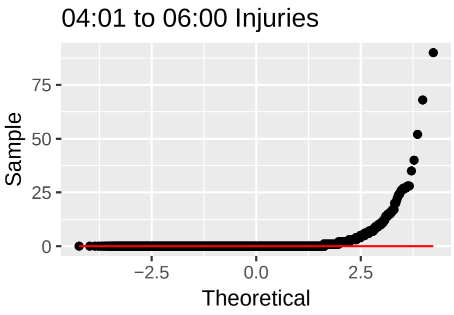
Felony Deaths

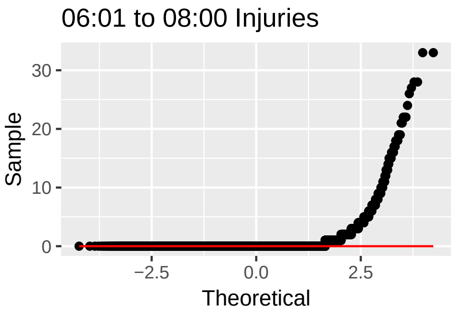
*Graphs are given in individual units of deaths or injuries.*

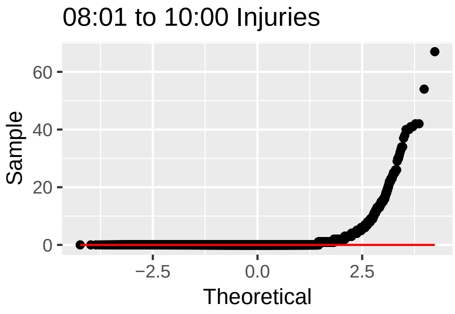
### Officer Injury and Fatality Probability Plots

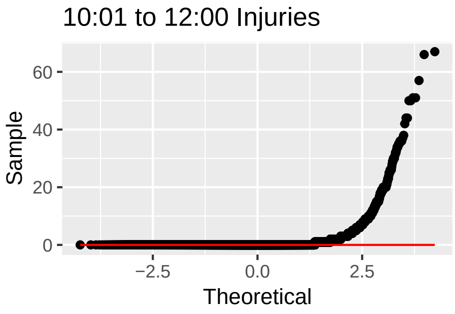

00:01 to 02:00 Injuries

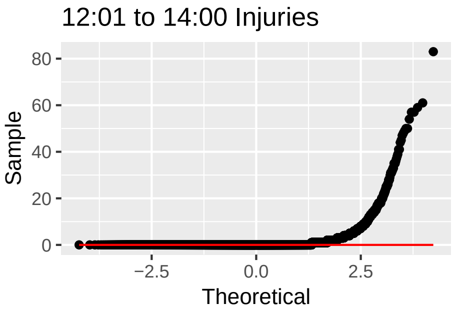

02:01 to 04:00 Injuries

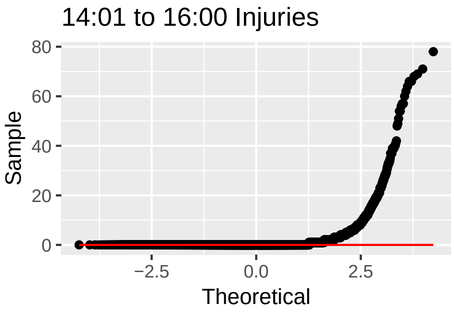

04:01 to 06:00 Injuries

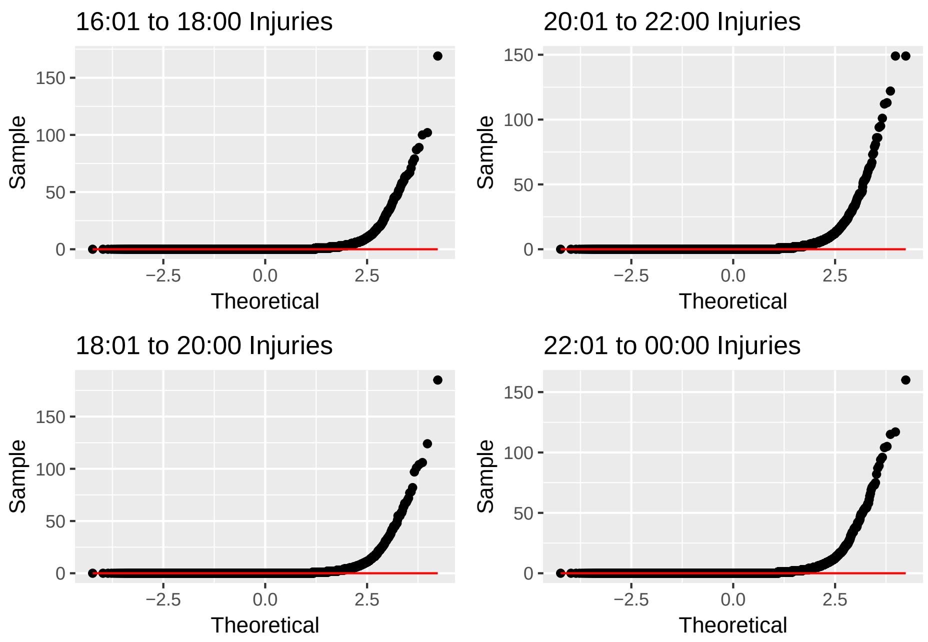

06:01 to 08:00 Injuries

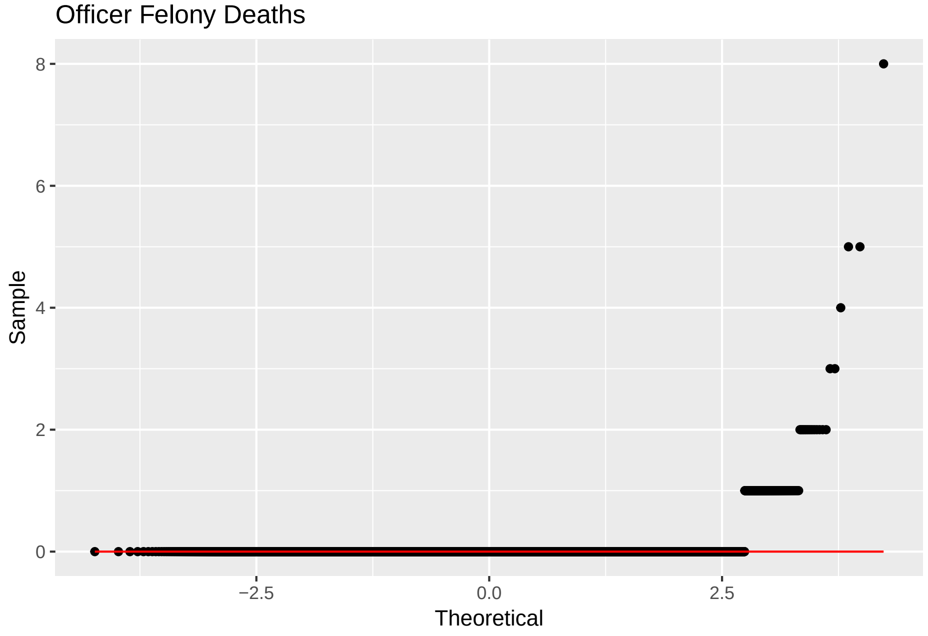

08:01 to 10:00 Injuries


10:01 to 12:00 Injuries


12:01 to 14:00 Injuries


14:01 to 16:00 Injuries

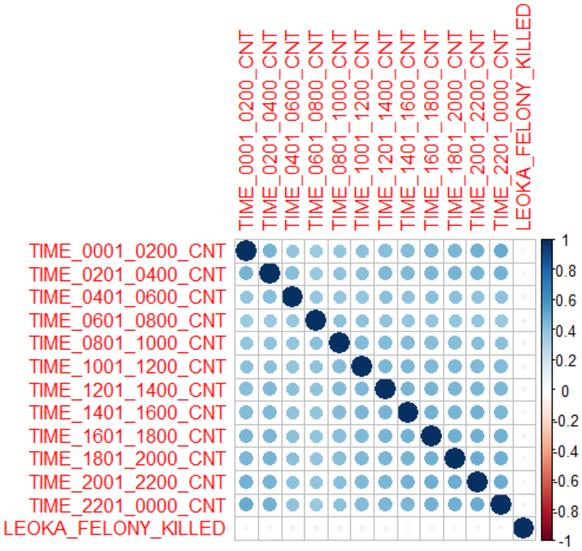




### Stat.desc Function (using a random sampling)

```
##                TIME_0001_0200_CNT TIME_0201_0400_CNT TIME_0401_0600_CNT
## median                       0.00               0.00               0.00
## mean                         0.46               0.28               0.13
## SE.mean                      0.03               0.03               0.01
## CI.mean.0.95                 0.06               0.05               0.02
## var                          4.86               3.37               0.75
## std.dev                      2.21               1.84               0.86
## coef.var                     4.81               6.55               6.86
## skewness                    13.03              25.20              17.15
## skew.2SE                   188.13             363.88             247.67
## kurtosis                 268.90             987.24             422.72
## kurt.2SE                1941.56            7128.38            3052.21
## normtest.W                 0.20               0.13               0.13
## normtest.p                 0.00               0.00               0.00
##              TIME_0601_0800_CNT TIME_0801_1000_CNT TIME_1001_1200_CNT
## median                     0.00               0.00               0.00
## mean                       0.12               0.20               0.25
## SE.mean                    0.01               0.02               0.02
## CI.mean.0.95               0.03               0.04               0.04
## var                        0.88               2.24               2.36
## std.dev                    0.94               1.50               1.54
## coef.var                   7.91               7.50               6.08
## skewness                  20.75              25.08              14.31
## skew.2SE                 299.55             362.07             206.68
## kurtosis                 578.48             918.91             285.26
## kurt.2SE                4176.92            6634.95            2059.75
## normtest.W                 0.10               0.11               0.15
## normtest.p                 0.00               0.00               0.00
##              TIME_1201_1400_CNT TIME_1401_1600_CNT TIME_1601_1800_CNT
## median                     0.00               0.00               0.00
## mean                       0.27               0.34               0.37
## SE.mean                    0.02               0.03               0.03
## CI.mean.0.95               0.04               0.06               0.07
## var                        2.41               4.42               5.51
## std.dev                    1.55               2.10               2.35
## coef.var                   5.75               6.21               6.29
## skewness                  15.83              18.07              17.94
## skew.2SE                 228.54             260.85             259.05
## kurtosis                 366.23             436.20             429.99
## kurt.2SE                2644.32            3149.56            3104.74
## normtest.W                 0.16               0.14               0.13
## normtest.p                 0.00               0.00               0.00
##              TIME_1801_2000_CNT TIME_2001_2200_CNT TIME_2201_0000_CNT
## median                     0.00               0.00               0.00
## mean                       0.42               0.46               0.45
## SE.mean                    0.03               0.04               0.03
## CI.mean.0.95               0.07               0.08               0.06
## var                        5.97               7.38               5.49
## std.dev                    2.44               2.72               2.34
## coef.var                   5.82               5.88               5.23
## skewness                  17.36              19.11              14.21
## skew.2SE                 250.67             275.84             205.11
## kurtosis                 418.73             527.16             280.17
## kurt.2SE                3023.42            3806.35            2022.94
## normtest.W                 0.15               0.14               0.18
## normtest.p                 0.00               0.00               0.00
##              LEOKA_FELONY_KILLED
## median                      0.00
## mean                        0.00
## SE.mean                     0.00
## CI.mean.0.95                0.00
## var                         0.00
## std.dev                     0.06
## coef.var                   18.73
## skewness                   20.29
## skew.2SE                  292.96
## kurtosis                  458.07
## kurt.2SE                 3307.49
## normtest.W                  0.03
## normtest.p                  0.00
```

All these graphs and the values returned in the stat.desc funciton indicate that the data is not normally distributed. There is significant right skew in every case. This is due to the reporting method by agency and they vary greatly in size. The Kenddall method was chosen for correlation analysis due to the high degree of ranking at the bottom of each graph.



While the circle graph shows some slight variance in correlation, no time frame stood out as significantly higher than the others.

## Machine Learning

For this section, I decided to determine if an effective model could be used to determine gender of the deceased from the police shooting data using the other values in the data frame. I used glmulti to obtain a model which is summarized here. Given the categorical nature of the data and the number of dimensions plotting would prove quite challenging.

```
##
## Call:
## glm(formula = gender ~ 1 + threat_level + age + threat_level:age +
##     signs_of_mental_illness:age + flee:age + race:age, family = binomial,
##     data = FatPShoot1416)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9304   0.1421   0.2627   0.3412   1.0489
##
## Coefficients:
##                                  Estimate Std. Error z value
## (Intercept)                      3.299572  0.478260   6.899
## threat_levelother               -2.158106  0.761722  -2.833
## threat_levelundetermined        -1.194454  3.420762  -0.349
## age                              0.723212 39.011466   0.019
## threat_levelother:age            0.055042  0.021445   2.567
## threat_levelundetermined:age     0.069135  0.115855   0.597
## age:signs_of_mental_illnessTrue -0.018808  0.006587  -2.855
## age:fleeFoot                     0.090813  0.032825   2.767
## age:fleeNot fleeing              0.019988  0.007947   2.515
## age:fleeOther                    0.027556  0.024789   1.112
## age:raceB                       -0.740823 39.011463  -0.019
## age:raceH                       -0.694615 39.011466  -0.018
## age:raceN                       -0.796400 39.011466  -0.020
## age:raceO                       -0.758276 39.011468  -0.019
## age:raceW                       -0.742366 39.011464  -0.019
##                                         Pr(>|z|)
## (Intercept)                     0.00000000000523 ***
## threat_levelother                        0.00461 **
## threat_levelundetermined                 0.72696
## age                                      0.98521
## threat_levelother:age                    0.01027 *
## threat_levelundetermined:age             0.55069
## age:signs_of_mental_illnessTrue          0.00430 **
## age:fleeFoot                             0.00566 **
## age:fleeNot fleeing                      0.01190 *
## age:fleeOther                            0.26630
## age:raceB                                0.98485
## age:raceH                                0.98579
## age:raceN                                0.98371
## age:raceO                                0.98449
## age:raceW                                0.98482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 644.69  on 1827  degrees of freedom
## Residual deviance: 584.57  on 1813  degrees of freedom
##   (130 observations deleted due to missingness)
## AIC: 614.57
##
## Number of Fisher Scoring iterations: 17
```

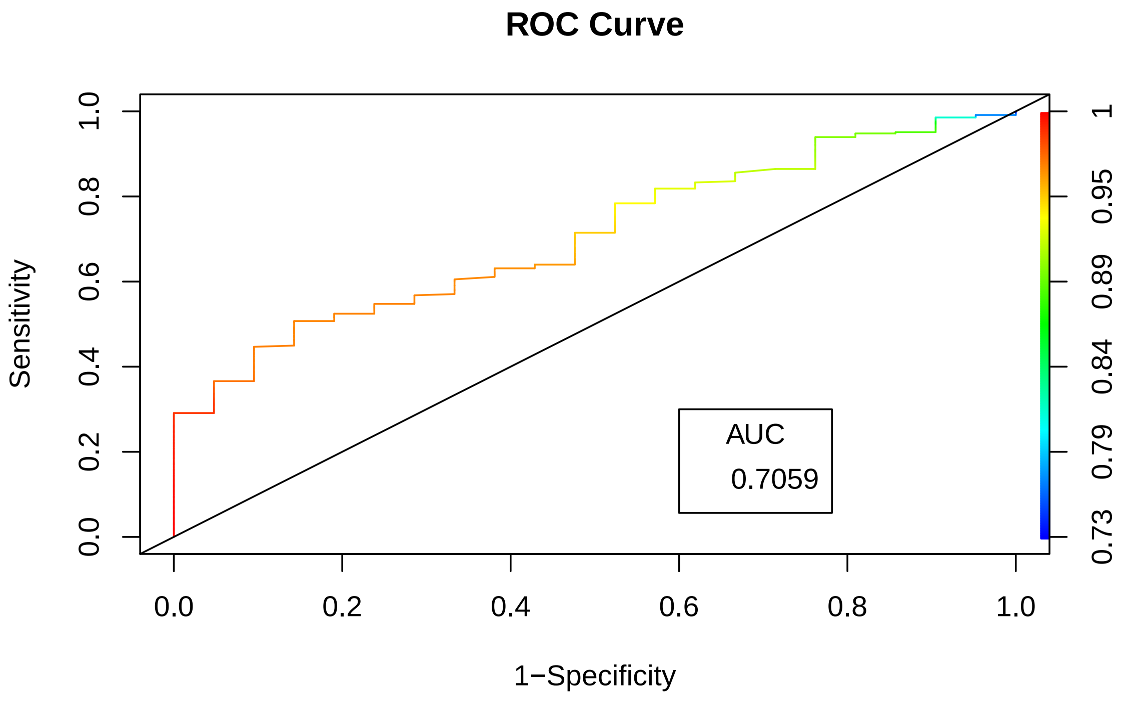
This model resulted in only 89% accuracy. The model was much more successful at determining male individuals, but the data was also strongly skewed male. This can be seen in the matrix shown below. Accuracy was determined in two ways. First the entire data set was used and then the data was split into an 80 percent training model and accuracy was determined against the remaining 20. The same accuracy value was returned for each method.

The following matrix shows the model against the entire data set.

```
##            ACTUAL
## PREDICTIONS    F    M
##       FALSE    4  126
##       M       78 1750
```

This result was computed using a training model and the confusionmatrix() function from the caret package.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   F   M
##          F   3  21
##          M  21 347
##
##                Accuracy : 0.8929
##                  95% CI : (0.8579, 0.9217)
##     No Information Rate : 0.9388
##     P-Value [Acc > NIR] : 0.9998
##
##                   Kappa : 0.0679
##  Mcnemar's Test P-Value : 1.0000
##
##               Precision : 0.125000
##                  Recall : 0.125000
##                      F1 : 0.125000
##              Prevalence : 0.061224
##          Detection Rate : 0.007653
##    Detection Prevalence : 0.061224
##       Balanced Accuracy : 0.533967
##
##        'Positive' Class : F
##
```



Since this test was done to determine if there were any variables worth exploring in relation to determining gender instead of requiring an effective model for answering key questions, I concluded the experiments at this point as both metrics of accuracy and area under the curve showed the model as less effective than desired.

## Conclusions and Reporting

This research started with a few questions around criminal justice data and how it might impact my son's proposed career in law enforcement. I have an interesting update on that front, but I'll reserve that information until after I report on the relevant questions. I opened my research up to further questions and insights as I explored the data and obtained another question from the class around immigration status. After exploring my initial questions on safety, I focused on issues of race, gender and residency. Key to this process were method found in the data science process. From data acquisition to data cleansing to exploratory data analysis to modelling and finally to the reporting which I am about to conclude.

I will take a father's liberty and first report on the data that mattered most to me. Looking at this data, I would indeed encourage my son to move to a safer area of the country like the North East. I would encourage him to avoid small cities with populations under 100,000. I did find that surprising and expected large cities to prove to be the most dangerous. The data on safer shifts did not show a clear shift with a strong correlation to danger. I found that surprising as I would have thought the night shift more dangerous based anecdotal media coverage. There were correlations to safer hours, but the correlation was subtle. Fortunately for me and before I shared this data with him, my son took me to the side a few weeks ago and told me that he had decided to switch majors and become an engineer. I see the irony given all the work, but given how much I have learned, and that the data proved more interesting than I would have thought without that context, I am content reporting his decision made without the information obtained from my project.

Both genders showed positive correlations to the total number of arrests, but the data showed a stronger correlation for males. That aligned with my preconceived notions and early graphing in the data exploration. Caucasians proved to have the strongest correlation to total arrests in addition to having the strongest correlation to police shootings. Given the more frequent interactions through arrests that would seem to make sense, but a more exhaustive exploration of the data accounting for population adjustments would provide clearer analysis and was beyond the scope of my work. And while it took more wrangling of the data than I would have liked, residents had a stronger correlation to arrests than non-residents and correlation for Hispanic non-residents proved even weaker still. Given all the media coverage and fascination that politicians have on this topic lends credence to it being more hype and conjecture from anecdotal evidence than from true statistical evaluations.

The answers above addressed the problem statement of this work. I gained a clearer understanding of the dangers law enforcement officers face, and I saw correlations to crime that did not match a lot of commonly held beliefs. I dispelled some notions that I went into this with and while it ultimately did not help me give guidance to my son, I learned a lot and felt better armed by statistical information along the way. In doing so I felt some relief about his previous career selection. And knowing my son, I could easily be revisiting this report after another semester.

My work was far from exhaustive given how inexperienced I am in this process, and given the scope of the data, a lot more analysis could provide fascinating insights into many other aspects of the data and further analysis on what I have started. Someone with a lot more experience with machine learning could bring more tools to dig deeper into many of my questions. Years could be spent pulling in supporting data and deeper analysis of the data. It could be summarized differently and complemented by supporting work on race, ethnicity and gender. Going through this learning process and coming up with my own conclusions using tools that I have learned from this class, evokes a strong sense of accomplishment. I am eager to learn more and build on these tools in future courses in the program.